行政院國家科學委員會專題研究計畫　成果報告

強化可擴充性叢集式網站伺服器的效能

計畫類別：個別型計畫
計畫編號：NSC91-2213-E-032-028-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：淡江大學資訊工程研究所

計畫主持人：陳伯榮

計畫參與人員：楊士央、陳立勳、詹念怡

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 10 月 16 日

行政院國家科學委員會專題研究計畫成果報告

強化可擴充式叢集式網站伺服器效能

Engineering Scalable Cluster-based Web Server for Performance

計畫編號：NSC 91－2213－E－032－028

執行期限：91 年 8 月 1 日至 92 年 7 月 31 日

主持人：陳伯榮　淡江大學資訊工程學系

共同主持人：林丕靜　淡江大學資訊工程學系

一、中文摘要

近年來，網際網路及網頁開發技術受到極大的關注及重視，設計者都希望透過叢集式網頁伺服器(Cluster-based web server)執行網頁模式應用程式(Web-based application)以提供具備可擴充性(Scalable)及高服務能力(highly available)的網頁服務。而效能往往是決定叢集式網頁伺服器是否成功的關鍵。

改善效能的方法很多，即便是網頁設計的改良也可有效提升效能，本計劃藉由影像傳輸及網頁挖掘(Web Mining)等相關領域技術來協助提昇叢集式伺服器效能。

我們探討如何利用漸進式影像傳輸技術(Progressive Image Transmission)透過漸進式的影像顯示方式，使得網頁的使用者可以在影像資料尚未完成傳輸前，即可透過較低解析的影像內容，預先了解內容，進而決定是否繼續等待資料傳輸或繼續進行其他工作。我們利用網頁使用者特性分析(Web Usage Analysis) 的觀念與方法去找出使用者使用網站的習性及對網頁資料的可能需求，以提供網站的管理人員或管理系統可以在必要時調整網站的內容或資料儲存位置配置。

關鍵詞：叢集式網頁伺服器，網頁模式應用程式，可擴充性，高服務能力，網頁挖掘，漸進式影像傳輸技術，網頁使用者特性分析

Abstract

Within a short time, the internet and WWW have become surpassing all other technological developments. Cluster-based Server is often used to run web-based applications to provide scalable and highly available web sites. Performance is one of the main problems in building cluster-based web server.

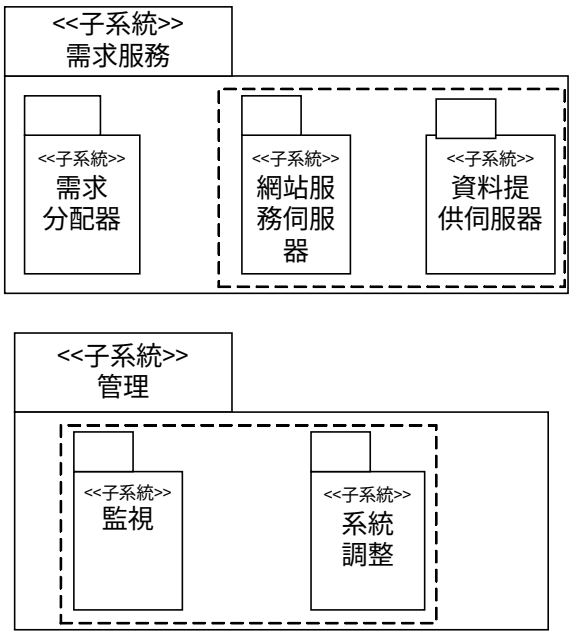
There are a number of techniques, which can be applied to improve performance; even web page design can have significant impact in performance. We study and then employ techniques from varied research area, such as image communication and web mining to engineer cluster-based web server for performance.

We study the progressive image transmission methods in image communication to transfer an image gradually, so that the viewer can see an approximated image in its whole to see if the image fits his/her need, without the need to wait for all the data to be received. We study the web usage analysis concepts and methods in web mining to understand how users view the data and how they actually use the site, so that the web master can reallocate or/and replicate the file adaptively, if necessary.

Keywords: Cluster-based server, web-based application, Scalable, highly available, web mining, web engineering, progress image transmission, web usage analysis

二、緣由與目的

我們已於[1]中提出一個叢集式網站伺服器系統的架構，其主要分為需求服務及管理兩個子系統，架構圖如下：



圖一：叢集式網站伺服器架構圖

為了提昇叢集式網站伺服器的效能，我們探討兩個主題包括：

1. 利用影像傳輸技術(Image Communication)中的漸進式影像傳輸(Progressive Image Transmission) [2][3]技術提昇服務效能及節省不必要傳輸頻寬：我們都知道圖形檔案佔據了網際網路傳輸的絕大部分頻寬，而網站的讀者往往直到檔案完成傳輸才知道這個圖檔是否正如讀者所需。這樣的過程不但浪費頻寬也降低使用者的效率。所幸，網際網路上所最常使用的兩種圖形格式：JPEG 及 GIF，都提供漸進式的解析模式(progressive coding option)，透過漸進式影像傳輸技術，使得網頁的使用者可以在影像資料尚未完成傳輸前，即可透過較低解析的影像內容，預先了解內容，進而決定是否繼續等待資料傳輸或繼續進行其他工作。

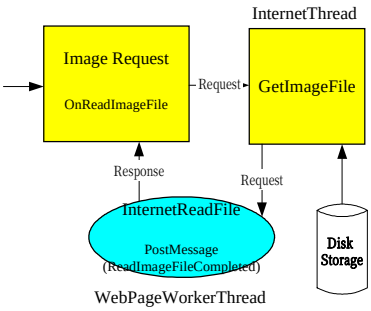
2. 利用網頁挖掘(Web Mining)分析使用者需求以協助叢集式伺服器進行負載調整或檔案配置[4][5]：Chen[6]最早提出利用資料挖掘(Data Mining)的方法來分析網頁紀錄資料(Web log data)，Cooley[7]提供由網頁紀錄資料找出使用者導覽區段(User Navigation Session)的方法與步驟。

我們將於下一節進一步說明這兩個主題。

三、結果與討論

（一）漸進式影像傳輸技術

在需求服務子系統中的網站服務伺服器子系統內我們實作漸進式圖像傳輸的模組架構設計，如圖二；在需求服務子系統中的需求分配器子系統部分，我們應用 ISAPI （Internet Service Application Provide Interface）技術，當需求分配器接到使用者的圖像傳輸需求時，使用者需求會觸發 OnReadImageFile 事件，此時再依照使用者的圖像讀取需求位址至磁碟中取出圖像檔案。



圖二：漸進式圖像傳輸

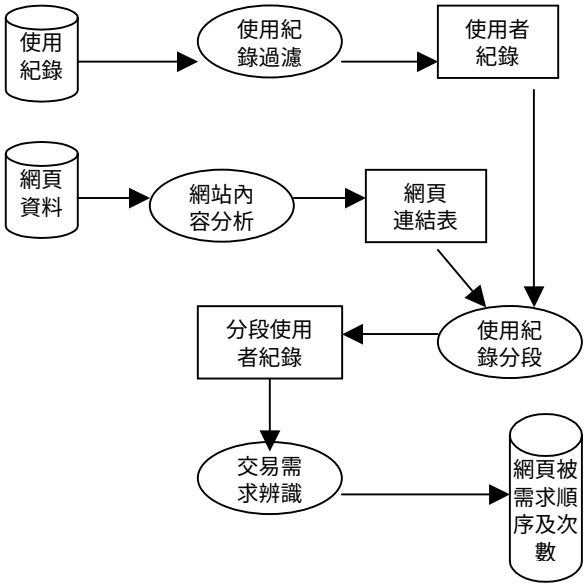
為了有效傳送及接收漸進式圖像，由於本模組旨在建立一有效且完整的機制，故而我們採用較為簡易的方式來處理漸進式圖像檔案，並且我們僅針對目前廣泛使用於 Internet 上之 JPEG 圖檔格式[8]來進行處理，此處將圖像的像點分割與處理原則分為五項，並簡略說明我們所採用之概念如下：

1. IDCT 對原始圖像檔案解碼，此時我們得到的是原圖的 Bitmap 圖像。
2. 將原圖像分割為 8 x 8 個區塊(Block)，共計 64 個像點，但此處整個圖像檔的傳輸是分為 8 次來傳輸完畢。
3. 依斜線方向分割像點。傳輸的像點個數會累次遞增；第一次為 1 個像點，第二次為 2 個像點，第三次為 3 個像點，…，第七次為 7 個像點。此處要注意的是：第八次傳輸的像點數則是前面七次的總合(共計 36 個像點)。
4. 取出各層像點，並重新組合。
5. 利用 Huffman Encoding 重新編碼。

（二）網頁挖掘技術

在圖一管理子系統中之監視子系統，我們分析使用者的紀錄以了解使用者對網站伺服器真正的需求。使用者需求對網站服務伺服器所帶來的負載成為我們了解的重心。

為了完成這個目標，我們利用網站使用挖掘(web usage mining)的觀念及方法來找出網站伺服器中使用者對於網站的使用習性及對網頁資料的需求。我們提出簡化並整合 Cooley[7]與 Chen[6]的步驟，在[4]中依照網站內容分析、使用紀錄過濾、使用紀錄分段及交易需求辨識等四個步驟說明我們的做法如圖三。



圖三： 簡化與整合的方法與步驟

進一步，在[5] 中為了適應於多變的使用者瀏覽行為，我們提出了使用一個模糊概念的完整預測系統架構。所謂模糊預測的概念，即選擇所有滿足相似度門檻值的預測模式；不同於單一預測概念，即選擇與使用者行為模式是最類似的一個預測模式。因此我們提出以趨勢相似度作為預測的基礎。此外，也試著對我們提出的系統所預測的網頁進行排序的動作；因為，多數的研究只提出哪些頁面是使用者會在未來使用到，但卻不關心這些頁面將被瀏覽到的順序為何，然而，我們卻覺得如果可以找到這些頁面的順序性，將可以提供我們一些有用及有趣的資訊。

我們提出兩個評估預測模型的方式：

頁面的正確率以及順序的正確率，用這兩個評量方式評量預測模型的準確度，最後在實驗階段，輔以 Hit Ratio 的值，以評估預測系統的效能。我們的預測系統分為兩個階段：

第一個階段：預測模型建立階段(Prediction Model Constructing Model)，將經由 Log Database 做資料前處理(Data Preprocessing)後，整理出的使用者行為區段(User's Behavior Session)，利用資料挖掘找到預測模式集。

在第二個階段：使用者未來瀏覽行為預測階段(Predicting Phase)，當新的使用者進入，我們依使用者瀏覽網站的瀏覽行為和預測模式中的比對模式(Match Pattern)去做趨勢相似度(Trend Similarity)的計算可分為下列兩種情形：

(1) 若有符合相似度門檻值的預測模式時，則挑選出符合門檻值以上的預測模式，做為可預測此一新使用者的參考模式，並計算出與每一個參考模式的趨勢相似度，然後，再利用預測器(Predictor)做出最後的預測，最後，我們的預測結果將得到一組有順序性的頁面，所以，我們的預測系統除了提供主機哪些網頁即將被需要存取到外，甚至，可以提供使用者需要的先後順序。

(2) 若沒有符合相似度門檻值的預測模式時，我們則將此一使用者的瀏覽行為儲存至暫存行為模式庫(Temporal Pattern Database)中，等到具有一定可信度時，我們將整合出新的預測模式，加入預測模型中。藉由這樣的回饋機制(Feedback)，不但可以保留預測系統的彈性，更可以適應多變的使用者行為，以提高整體預測的效能與準確度。

為了有效改善叢集式網站伺服器系統效能，我們在[3]中利用漸進式影像傳輸技術透過漸進式的影像顯示方式，使得網頁的使用者可以在影像資料尚未完成傳輸前，即可透過較低解析的影像內容，預先了解內容，進而決定是否繼續等待資料傳輸或繼續進行其他工作。以進一步提昇系統服務的效能。

另外，由於網路使用者的瀏覽行為是非常多變的，所以，一個好的預測系統除了預測準確度要高之外，也必須適應多變的使用者行為。而我們提出預測系統，以模糊理論的預測概念去適應多變的網路使用者瀏覽行為，由測試的結果觀察出這樣的預測概念較比對單一模式做預測為佳；此外，我們也由測試結果得知以趨勢相似度做預測比以瀏覽頁面的相似度做預測來的好。而在我們提出的預測系統中，也加入回饋機制以保留預測系統的彈性與可調適性。

在未來的研究中，我們希望能夠再提升預測系統的準確度，以及順序的正確度。並將回饋機制再加以改進，使回饋機制更為完整，

四、計劃成果自評

五、參考文獻

[1] 陳伯榮、楊士央、李文禮、陳立勳，”叢集式網站伺服器架構”二十一世紀數位生活與網際網路科技研討會，成功大學，民國九十年五月, 國科會 NSC 90-2213-E-032-014 研究計劃。

[2] 陳伯榮、楊士央、陳立勳，”整合 ISAPI 技術強化叢集式網站伺服器系統效能”，二○○三數位生活與網際網路科技研討會，2003 年 9 月

[3] 陳立勳,” 整合 ISAPI 技術強化叢集式網站伺服系統效能” 私立淡江大學資訊工程研究所碩士論文,2003,6.

[4] 陳伯榮、楊士央、陳立勳、詹念怡，”如何應用網頁使用挖掘方法來提高網頁伺服器效能與能力”，二○○二數位生活與網際網路科技研討會，2002 年 6 月

[5] 詹念怡，” ” 私立淡江大學資訊工程研究所碩士論文,2003,6.

[6] M. S. Chen, J. S. Park, and P. S. Yu. “Data Mining for Path Traversal Patters in a Web Environment,” Proc. 16th Int’l Conf. On Distributed Computing Systems, pp. 385-392, 1996.

[7] R. Cooley, B. Mobasher, and J. Srivastava, “Data preparation for mining world wide web browsing patterns, ”Knowledge end Information Systems, 1(1): 5-32, Feburary 1999.

[8] Gregory K. Wallace, “The JPEG Still Picture Compression Standard”, Submitted in December 1991 for publication in IEEE Transactions on Consumer Electronics

附件：封面格式

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※※※

※　　　強化可擴充式叢集式網站伺服器效能

※

※

※

※

※※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：□個別型計畫　　□整合型計畫

計畫編號：NSC　91－2213－E－032－028－

執行期間：　　91 年　8 月　1 日至　92 年　7 月　31 日

計畫主持人：陳伯榮

共同主持人：林丕靜

本成果報告包括以下應繳交之附件：

□赴國外出差或研習心得報告一份

□赴大陸地區出差或研習心得報告一份

□出席國際學術會議心得報告及發表之論文各一份

□國際合作研究計畫國外研究報告書一份

執行單位：淡江大學

中　華　民　國　92　年　10 月　　16 日